BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35206
doi: 10.3897/biss.3.35206

OPEN ACCESS

Conference Abstract

Wikidata as a linked-data hub for Biodiversity data

Andra Waagmeester[‡], Lynn Schriml[§], Andrew Su[|]

‡ Micelio, Ekeren (Antwerpen), Belgium
§ Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, United States of America
| Scripps Research Institute, La Jolla, United States of America

Corresponding author: Andra Waagmeester (andra@micelio.be)

Received: 05 Apr 2019 | Published: 18 Jun 2019

Citation: Waagmeester A, Schriml L, Su A (2019) Wikidata as a linked-data hub for Biodiversity data . Biodiversity Information Science and Standards 3: e35206. https://doi.org/10.3897/biss.3.35206

Abstract

Wikidata (http://www.wikidata.org) is the linked database of the Wikimedia Foundation. Like its sister project Wikipedia it is open to humans and machines. Initially primarily intended as a central repository of structured data for the approximately 200 language versions of Wikipedia, Wikidata currently also serves many other use cases. It is an open, Semantic Web-compatible database that anyone can edit.

Here, we present the Gene Wiki initiative. In 2008, this project started by creating Wikipedia articles for all human genes (Huss et al. 2008). These articles were enriched with structured information on these genes as tables (called infoboxes). With the onset of Wikidata in 2012, the project diverted its attention from the infoboxes and since we have been enriching Wikidata with structured knowledge from public scientific resources on gene, proteins, diseases and compounds (Burgstaller-Muehlbacher et al. 2016).

This structured information is added to Wikidata, while active links to the primary source are maintained. Adding a new resource to Wikidata is a community-driven process that starts with modelling the subjects of the resource under scrutiny. This involves seeking commonalities with similar concepts in Wikidata and, if none are found, new are created. This process mostly happens in a collaboratively-edited document (i.e. GDocs), where different graphical networks are drawn to reflect the data being modelled and its embedding in Wikidata. Once consensus has been reached, the model typically exists in a human-readable document. To allow future validations of these models on existing data, it

© Waagmeester A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

is converted in a machine-readable Shape Expression (ShEx) (Anonymous 2019, Waagmeester et al. 2017). Shape Expressions schema language can be consumed and produced by humans and machines and is useful in model development, legacy review or as formal documentation.

Once a semantic data model (as Shape Expression) is found, i.e. community consensus is reached, a bot is developed to convert the knowledge from the primary source, into the Wikidata model. While Wikidata is linked data (part of the semantic web), many life science resources are not. On the contrary, many distinct file formats or API output formats are used to present life-science knowledge. To convert between these different formats, bots need to be developed that are able to parse the different resources and serialize into wikidata. We have developed a software library in the Python programming language, which we use to build these bots. Once created, these bots run regularly to keep Wikidata up-to-date with knowledge on genes, proteins, diseases and drugs.

Having scientific knowledge represented in Wikidata comes with benefits. First, having research data on Wikidata increases its sustainability. When research projects end, their findings now remain on an independently funded infrastructure. Having someone else dealing with an infrastructure for a data commons also relieves the research community of having to do it themselves, leading to more time to focus on doing research

As a generic public data commons, Wikidata allows public scrutiny and rapid integration with other domains. Inconsistencies or disagreement between resources become more visible, due to the unified data models and interfaces.

The latter we leverage as a feature in our bots. One of our core resources is, for example, the Disease Ontology (Schriml et al. 2018). This ontology on human diseases is continuously updated by its curation team. 2 times per month, updates are then synchronised with Wikidata. If inconsistencies and disagreement with other resources surface, they are logged and shared with the curation team of the Disease Ontology. Hence, we have created a bi-directional update cycle, improving both the Disease Ontology and Wikidata.

Although our bots focus on molecular biology, our approaches are generic in onset that we are confident a similar approach can work in biodiversity informatics.

Keywords

wikidata, wikipedia, molecular biology, crowd-sourcing

Presenting author

Andra Waagmeester

Funding program

National Institutes of General Medical Sciences R01GM089820

References

- Anonymous (2019) SHEX - Shape Expressions: Validation, traversal and transformation of RDF graphs. http://shex.io. Accessed on: 2019-12-04.
- Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, Naik C, Pavlidis P, Schriml L, Good BM, Su AI (2016) Wikidata as a semantic framework for the Gene Wiki initiative. Database : the journal of biological databases and curation 2016 https://doi.org/10.1093/database/baw015
- Huss JW, Orozco C, Goodale J, Wu C, Batalov S, Vickers TJ, Valafar F, Su AI (2008) A gene Wiki for community annotation of gene function. PLoS biology 6 (7): e175. https://doi.org/10.1371/journal.pbio.0060175
- Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C (2018) Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Research 47 https://doi.org/10.1093/nar/gky1032
- Waagmeester A, Prud'Hommeaux E, Mitraka E, Stupp G, Queralt-Rosinach N, Burgstaller-Muehlbacher S (2017) Using Shape Expressions (ShEx) to model, validate and curate Wikidata. figshare. Poster